Airbnb Analysis & Price Prediction

Prepared by: Saleh Abdallah

Date: Feb 2, 2025

Introduction

Objective: To analyze Airbnb dataset in NYC and build a predictive model for price estimation.

Approach:

Data cleaning and preprocessing

Exploratory Data Analysis (EDA)

Model Training & Evaluation

Dataset Overview

Dataset: Airbnb listings from NYC

Key Features:

Price (Target variable)

Neighbourhood Group

Room Type

Number of Reviews

Availability_365

Preprocessing Steps:

Handling missing values

Outlier detection and removal

Feature Engineering

Exploratory Data Analysis (EDA)

Visualizations & Findings:

Number of reviews & availability distribution: Skewed right

Price Distribution: Presence of extreme values

Correlation Matrix: Weak correlation between price, number of reviews, and availability

Price variation across neighborhood groups and room types

Key Insights:

Manhattan and Brooklyn have the highest prices

Manhattan has the highest prices across all room types

Entire homes/apartments are priced the highest, followed by private rooms and shared rooms

Staten Island and Bronx have the higher availability

Shared room have the highest availability compared to tother room types

Feature Engineering & Model Building

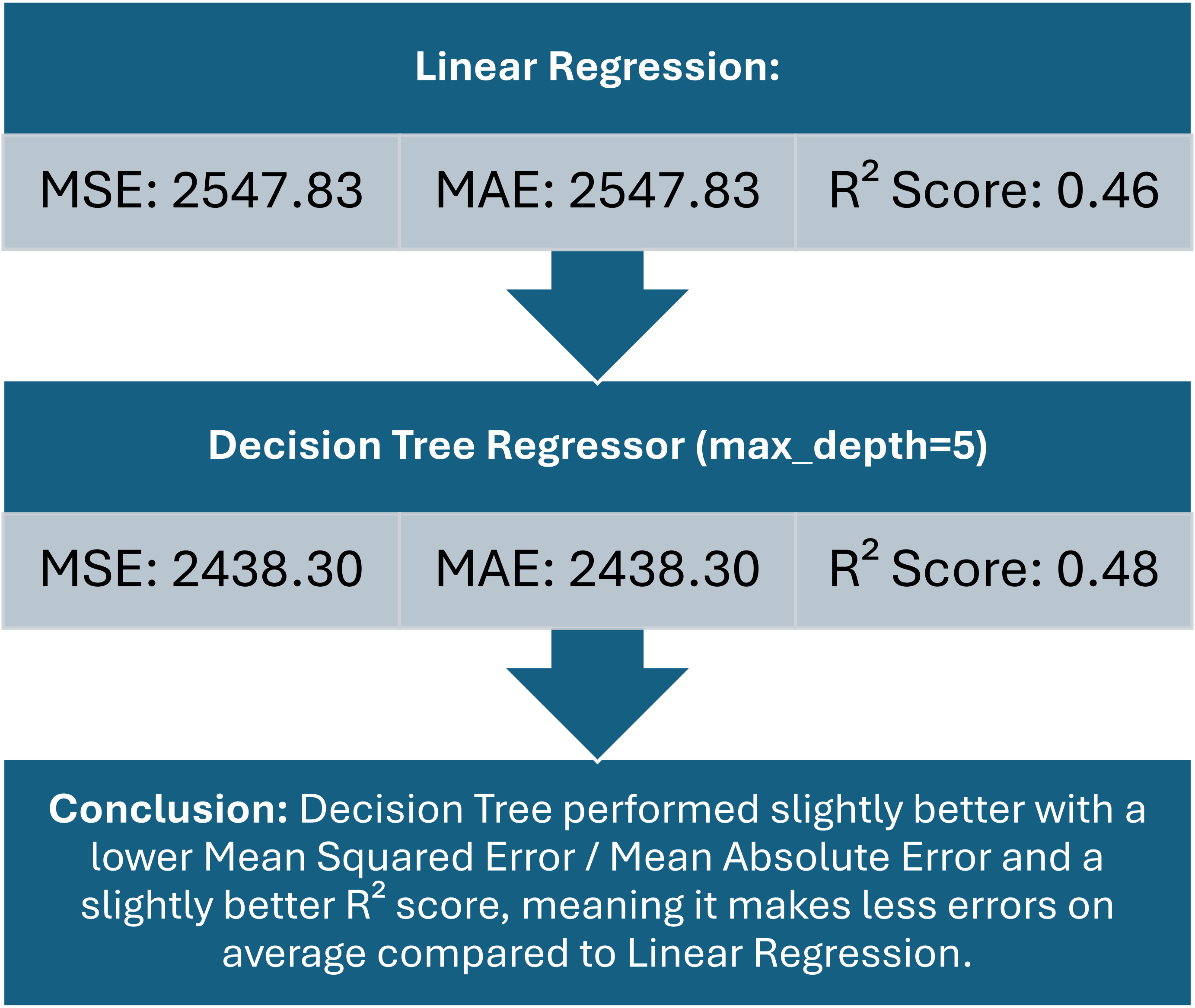
Feature engineering:

- Outlier removal using IQR method
- Feature selection significantly impacted model performance
- Including categorical features improved predictive accuracy

Models used:

- Linear Regression
- Decision Tree Regressor

Model Performance Comparison



Linear Regression:

MSE: 2547.83	MAE: 2547.83	R^2 Score: 0.46

Decision Tree Regressor (max_depth=5)

MSE: 2438.30	MAE: 2438.30	R^2 Score: 0.48

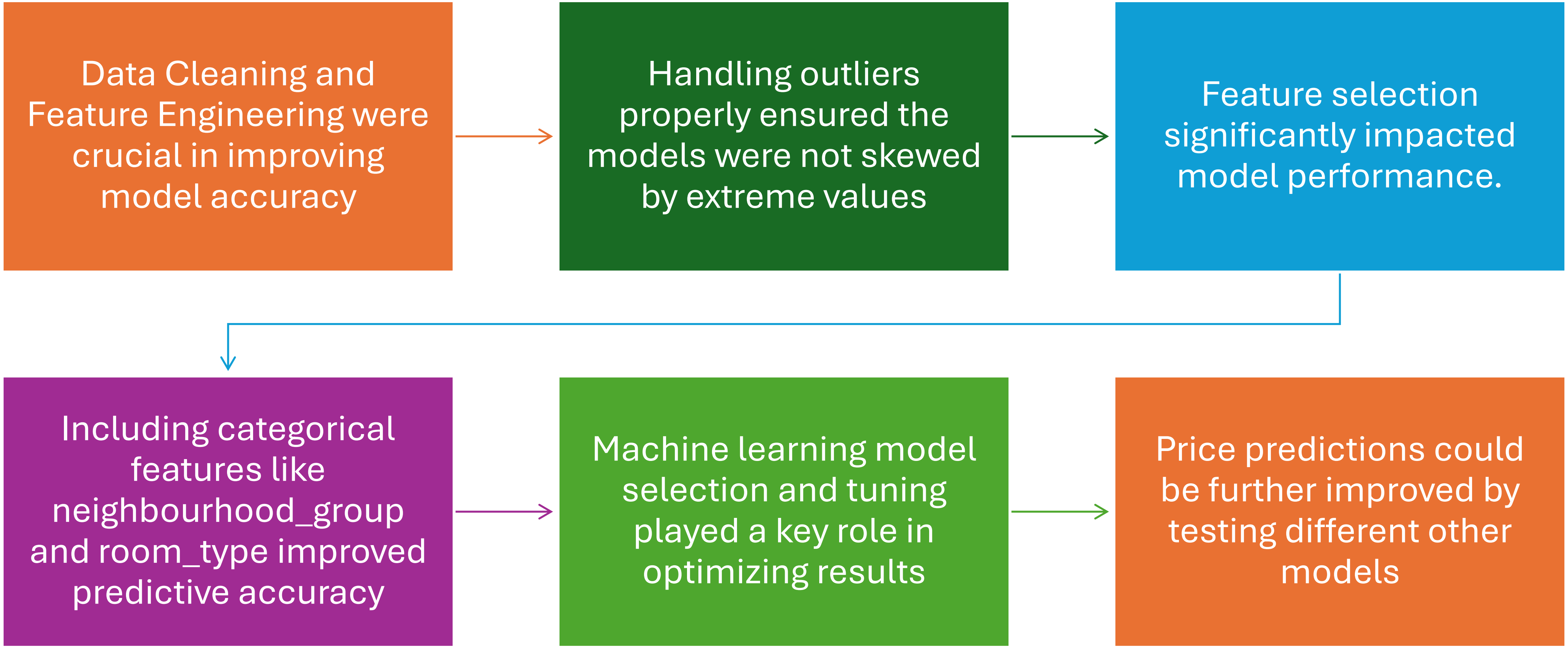
Conclusion: Decision Tree performed slightly better with a lower Mean Squared Error / Mean Absolute Error and a slightly better R^2 score, meaning it makes less errors on average compared to Linear Regression.

Key Findings & Insights

Increasing the decision tree depth improved the performance by reducing MSE/MAE and increasing the R^2 score

Decision Tree model slightly outperformed the Linear Regression model

Reflections and Learnings



Data Cleaning and Feature Engineering were crucial in improving model accuracy

Handling outliers properly ensured the models were not skewed by extreme values

Feature selection significantly impacted model performance.

Including categorical features like neighbourhood_group and room_type improved predictive accuracy

Machine learning model selection and tuning played a key role in optimizing results

Price predictions could be further improved by testing different other models